Reviewer #1

Comment #1-0. This manuscript presents a framework for generating fractional winter wheat maps at 1 km resolution across 53 countries for the period 2001–2020, using MODIS surface reflectance data combined with Random Forest regression and Gaussian Mixture Models (GMM). The topic is relevant to ESSD and addresses a genuine need for long-term, large-scale crop distribution data. The overall ambition—producing the first global winter wheat fraction time series at 1 km—is commendable.

However, the manuscript in its current form suffers from serious deficiencies in conceptual clarity, methodological rigor, and depth of analysis. The approach is almost entirely empirical/statistical, with insufficient physical or mechanistic grounding. Many core concepts are introduced without adequate definition or justification. The validation strategy has potential circularity issues. I recommend major revision, with the specific concerns detailed below.

Response: Thank you very much for your positive comments. We have carefully revised the manuscript in response to your feedback and provided a detailed, point-by-point response below.

Major Concerns

Comment #1-1. Lack of Physical/Mechanistic Foundation

The entire framework is built on a cascade of statistical models: Random Forest regression → Gaussian Mixture Model fitting → Random Forest threshold prediction. At no point does the paper ground any of this in crop phenology, radiative transfer physics, or the spectral properties of wheat canopies. This is a fundamental weakness for a paper that claims to provide a data product for the Earth system science community. Specific issues:

- Why do Gaussian parameters carry physical meaning? The paper states (L245–246) that the Gaussian model shows clear physical meaning regarding the dominant crop types. This claim is never substantiated. What is the biophysical interpretation of the mean (μ), standard deviation (σ), and amplitude (A) of a Gaussian fitted to the histogram of predicted wheat fractions within a 100 km grid? How do these parameters relate to planting intensity, field size, crop variety, or phenological stage? Without this link, the GMM is merely a curve-fitting exercise.

Response: thank you. We revised the manuscript to detail different characteristics of the winter wheat fraction distribution within each 100 km grid. First, the mean (μ) and amplitude (A) of a Gaussian model describe regional planting intensity and the spatial concentration of winter wheat, while the standard deviation (σ) reflects spatial heterogeneity associated with field fragmentation and crop mosaics. We also included relevant references to support these interpretations and clarified that Gaussian parameters are employed to describe regional planting patterns.

"The mean (μ) and amplitude (A) of the gaussian curve reflect the regional planting intensity and the spatial concentration of winter wheat within each 100 km grid, in which higher values of both parameters generally indicate a more dominant distribution of winter wheat in the regional cropping structure (Skakun, 2017). The standard deviation (σ) characterizes the spatial heterogeneity of winter wheat distribution, where larger values are associated with greater field fragmentation and a mosaic pattern of different crops (Liu, 2020; Masiliūnas, 2021)." (page 8, line 192-196)

References:

Liu, L., Xiao, X., Qin, Y., Wang, J., Xu, X., Hu, Y., and Qiao, Z.: Mapping cropping intensity in China using time series landsat and sentinel-2 images and Google earth engine, Remote Sens. Environ., 239, 111624, https://doi.org/10.1016/j.rse.2019.111624, 2020.

Masiliūnas, D., Tsendbazar, N.-E., Herold, M., Lesiv, M., Buchhorn, M., and Verbesselt, J.: Global land characterisation using land cover fractions at 100 m resolution, Remote Sens. Environ., 259, 112409, https://doi.org/10.1016/j.rse.2021.112409, 2021.

Skakun, S., Franch, B., Vermote, E., Roger, J.-C., Becker-Reshef, I., Justice, C., and Kussul, N.: Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a gaussian mixture model, Remote Sens. Environ., 195, 244–258, https://doi.org/10.1016/j.rse.2017.04.026, 2017.

- Why should a bimodal fraction distribution imply two crop types? The paper uses biGaussian fitting to discriminate wheat from non-wheat signals. But a bimodal distribution of RF-predicted fractions could arise from many factors: systematic RF bias, landscape fragmentation, topographic effects, atmospheric correction residuals, or artifacts of the MODIS compositing. The authors need to demonstrate—not assume—that the modes correspond to distinct crop types.

Response: thank you. Sorry, you might misunderstand our strategy. We applied the single- and bi-Gaussian distributions to characterize variations in winter wheat planting structure within each grid rather than to identify specific crop types. For clarify, we explained this issue in our revised manuscript.

"We applied the single and bi-Gaussian distributions to demonstrate variations in winter wheat planting structure within each grid, rather than to identify specific crop types." (page 8, line 203-205)

- Spectral confusion is mentioned but never quantified. Section 1 mentions spectral confusion within crop regions (L83–84), yet the method section provides no analysis of which crop types are most frequently confused with winter wheat, in which regions, or at which phenological stages. This is essential for users to understand the product's limitations.

Response: thank you. We determined the major crop types that may be confused with winter wheat across different regions, along with the specific phenological stages when this confusion is most probable. Specifically, we identified winter barley and winter rye in Europe, as well as winter rapeseed and winter garlic in China, as potential sources of spectral confusion. We also explained that confusion is usually more evident during the vegetative growth stage and might continue into the heading stage for some winter cereals. We also added the MODIS feature importance analysis of spectral-temporal regarding Random Forest regression model in the revised manuscript.

"For example, winter wheat exhibits spectral and phenological similarities to other winter cereals (e.g., winter barley and winter rye) in Europe, as well as winter rapeseed and winter garlic in China (Liu et al., 2024; Tian et al., 2024; Fu et al., 2025; Yang et al., 2025). Such confusion is typically more pronounced during the vegetative growth stage, especially from emergence and tillering to green-up and stem elongation, as crops display similar green-canopy traits and growth patterns over time (Akanbi et al., 2024). More specifically, spectral confusion between winter wheat and other winter cereals may persist until the heading stage, whereas winter garlic may exhibit spectral

and phenological patterns similar to those of winter wheat throughout much of the growing season." (page 2, line 60-66)

References:

Akanbi, O. D., Bhuvanagiri, D. C., Barcelos, E. I., Nihar, A., Gonzalez Hernandez, B., Yarus, J. M., and French, R. H.: Integrating multiscale geospatial analysis for monitoring crop growth, nutrient distribution, and hydrological dynamics in large-scale agricultural systems, J. Geovisualization Spatial Anal., 8, 9, https://doi.org/10.1007/s41651-023-00164-y, 2024.

Fu, Y., Chen, X., Song, C., Huang, X., Dong, J., Peng, Q., and Yuan, W.: High-resolution mapping of global winter-triticeae crops using a sample-free identification method, Earth Syst. Sci. Data, 17, 95–115, https://doi.org/10.5194/essd-17-95-2025, 2025.

Liu, M., He, W., and Zhang, H.: WPS: a whole phenology-based spectral feature selection method for mapping winter crop from time-series images, ISPRS J. Photogramm. Remote Sens., 210, 141–159, https://doi.org/10.1016/j.isprsjprs.2024.03.005, 2024.

Tian, H., Yang, M., Wu, F., Qin, Y., Zhang, X., Liu, J., and Yan, W.: Potentiality of landsat-9 for early-season mapping of winter garlic and winter wheat, Geo-spatial Inf. Sci., 27, 2199–2210, https://doi.org/10.1080/10095020.2024.2311868, 2024.

Yang, G., Li, X., Xiong, Y., He, M., Zhang, L., Jiang, C., Yao, X., Zhu, Y., Cao, W., and Cheng, T.: Annual winter wheat mapping for unveiling spatiotemporal patterns in China with a knowledge-guided approach and multi-source datasets, ISPRS J. Photogramm. Remote Sens., 225, 163–179, https://doi.org/10.1016/j.isprsjprs.2025.04.031, 2025.

- The paper never explains *why* MODIS bands can separate winter wheat fractions. Which spectral bands are most important in the Random Forest model? What phenological signals does MODIS capture for winter wheat specifically (e.g., the winter-spring green-up that distinguishes winter wheat from spring wheat)? The authors should report variable importance from the RF models and link these to known spectral-temporal signatures of winter wheat.

Response: thank you. We added a feature importance analysis of spectral-temporal regarding Random Forest regression model. First, B1, B2, and B7 are more important than the other MODIS bands, in whith B2 during the Green-up–jointing stage showing the highest importance, followed by B1 during the Green-up–jointing stage and B2 during the Heading–grain filling stage. On the other hand, the Green-up–jointing, Heading–grain filling, and Senescence–harvest stages generally contribute strongly to the mapping performance.

"4.1. Importance of input features derived from winter wheat regression model

Both the MODIS spectral and phenological features are crucial for mapping the spatiotemporal dynamics of winter wheat fractions (Fig. 5). This study measured the importance of input features to the mapping performance in the random forest regressor by summing the 5th, 50th, and 95th percentile importance scores for each band within each month, thereby better assessing the spectral-temporal characteristics of winter wheat mapping. In general, input features derived from the sixth to the tenth months of the winter wheat growing season (i.e., Green-up-jointing, Heading-grain filling, and Senescence-harvest stages) play vital roles in the mapping performance, suggesting the growing status in these stages is considerably different from other non-winter wheat types. Besides, input features such as B1, B2, and B7 are more crucial than B3, B4, B5, and B6. Specifically, B2 in

the Green-up-jointing stage is the most essential feature in our study, followed by B1 in the Green-up-jointing stage and B2 in the Heading-grain filling stage (i.e., red rectangles). These two bands are important to estimate the proportion of winter wheat, probably due to the rapid changes in canopy structure and chlorophyll content during the Green-up-jointing and Heading-grain filling stages (Fan et al., 2022; Li et al., 2024). In addition, B7 is essential to winter wheat classification as its relatively high importance after the Overwintering growing stage." (page 10, line 243-255)

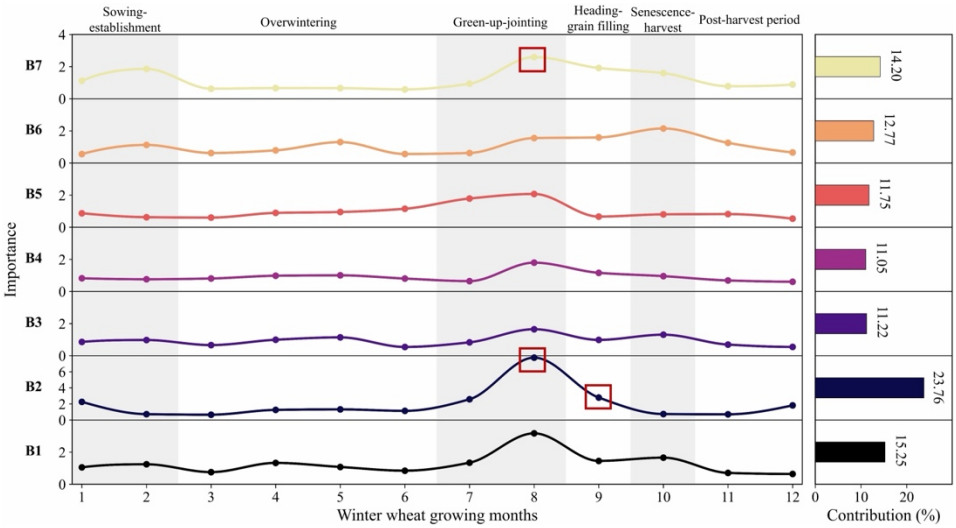


Fig 5. *The importance of spectral-temporal features in different months to the corn classification.*

References:

Fan, L., Xia, L., Yang, J., Sun, X., Wu, S., Qiu, B., Chen, J., Wu, W., and Yang, P.: A temporal-spatial deep learning network for winter wheat mapping using time-series sentinel-2 imagery, ISPRS J. Photogramm. Remote Sens., 214, 48–64, https://doi.org/10.1016/j.isprsjprs.2024.06.005, 2024.

Li, W., Weiss, M., Jay, S., Wei, S., Zhao, N., Comar, A., Lopez-Lozano, R., De Solan, B., Yu, Q., and Wu, W.: Daily monitoring of effective green area index and vegetation chlorophyll content from continuous acquisitions of a multi-band spectrometer over winter wheat, Remote Sens. Environ., 300, 113883, https://doi.org/10.1016/j.rse.2023.113883, 2024.

Comment #1-2. Validation Circularity and Independence

The validation strategy raises serious concerns about independence:

- Training and validation data come from the same sources. The fractional samples used for RF training are derived from the same CDL, EUCROPMAP, China winter wheat maps, and GlobalWheatYield4km products that are later used as reference data for validation. For instance, Section 4.3.2 compares the authors' results for China against aggregated maps from Dong et al. (2020)—the very dataset used to generate training samples. This is not independent validation.
- The FAO comparison (Section 4.3.1) is the only truly independent validation, but it is only at the national level, which masks pixel-scale errors. The reported R^2 of 0.81 at the country level does not justify the claim that the product is robust at the pixel scale.

Response: thank you for these two comments. We computed R and RMSE for each year by comparing our annual fractional mapping results with the ChinaWheat30L dataset at pixel scale.

We inserted a table in our revised manuscript.

"In addition, we also assessed the mapping area at the subregional and pixel scale between our mapping results and official statistics in several regions (e.g., China, the US, and EU countries)." (page 10, line 237-238)

"In addition, we compared our annual fractional mapping results with the ChinaWheat30L dataset from 2001 to 2020 to further assess mapping performance at the pixel scale. The comparison results of R and RMSE suggested excellent adaptation and effectiveness of the proposed winter wheat proportion mapping framework across years in our study (Table. S2)." (page 17, line 353-356)

Table S2. *Comparison of winter wheat fractional coverage at the pixel level in China between our findings and the ChinaWheat30L dataset.*

Year	R	RMSE	Year	R	RMSE
2001	0.72	13.08%	2011	0.76	12.57%
2002	0.73	11.58%	2012	0.73	11.95%
2003	0.71	12.40%	2013	0.75	13.47%
2004	0.75	12.73%	2014	0.72	10.61%
2005	0.72	10.99%	2015	0.74	15.11%
2006	0.74	11.00%	2016	0.77	10.92%
2007	0.73	11.05%	2017	0.78	12.77%
2008	0.77	10.96%	2018	0.76	11.88%
2009	0.72	11.60%	2019	0.79	10.25%
2010	0.74	12.18%	2020	0.78	16.42%

- The 70/30 train-test split at the grid level (Section 3.2) is insufficient to guarantee independence when grids are spatially autocorrelated. The authors must demonstrate that the training and validation grids are spatially independent, e.g., through a spatial block cross-validation scheme.

Response: thank you. We implemented a spatial block validation method, where each 100 km spatial block is assigned solely to either the training or validation dataset. The R^2 between the predicted and reference optimal thresholds decreased slightly from 0.972 to 0.926 after applying spatial block validation, indicating the threshold estimation remained stable after reducing spatial dependence between the training and validation data.

"In addition, we assigned entire 100km spatial blocks exclusively to either the training or validation dataset to reduce spatial dependence between the training and validation data. The R^2 value between the predicted and reference optimal thresholds was 0.926, indicating the threshold estimation performance remained stable after reducing spatial dependence between the training and validation grids." (page 13, line 297-300)

- No comparison with independent field survey data. Given the availability of in-situ crop type data (e.g., LUCAS for Europe, field surveys by agricultural agencies), the authors should include validation against ground-truth points that were not used in any form during model development.

Response: thank you. Our target variable is the winter wheat fraction within each 1 km × 1 km pixel, not a specific crop label at a point. Field survey data such as LUCAS typically offer crop details at exact points, but they do not directly indicate the proportion of winter wheat within a 1 km pixel. Obtaining accurate 1 km-scale winter wheat fractions requires dense and representative

field observations, which is difficult over the extensive global area and 20-year timeframe. Therefore, we validated our findings using independent crop data and agricultural statistics across various spatial scales.

Comment #1-3. Temporal Extrapolation Without Validation

The paper describes a temporal extension strategy (L176–178) where winter wheat samples from 2020 are used to train models for earlier years (2001–2019). This is a critical step that receives almost no attention:

- The assumption of temporal stationarity in wheat spectral signatures is never discussed, let alone tested. Crop varieties, planting practices, and climate conditions change over two decades. A sample of winter wheat in Kansas in 2020 may not be spectrally representative of winter wheat in 2010, let alone 2001.
- No temporal validation is performed for years before the reference data exist. For China, the reference product begins in 2016, yet results are mapped back to 2001. How do the authors know the model works for 2001–2015?
- The authors should at minimum: (a) use a subset of years with available reference data to test temporal transferability, (b) report separate accuracy metrics for early vs. late periods, and (c) discuss how changes in MODIS sensor performance (e.g., Terra degradation over 20 years) may affect results.

Response: thank you for these comments. We computed R^2 for each year by comparing our annual fractional mapping results with the ChinaWheat30L dataset. The model trained on data from 2016-2020 achieved mean R^2 values of 0.95 for 2001-2015 and 0.96 for 2016–2020, suggesting it has a strong ability to transfer across these time periods. In addition, we identified a relatively large systematic underestimation during 2001-2015 and discussed this limitation in the revised manuscript. In addition, we identified a relatively large systematic underestimation from 2001 to 2015 and addressed this limitation in the revised manuscript. We also expanded the discussion of temporal changes in winter wheat spectral and phenological characteristics and considered the potential effects of long-term MODIS sensor performance and product stability on multi-decadal mapping.

"The temporal dynamics of winter wheat area between our results and the ChinaWheat30L data show good performance regarding the temporal extension in our study (Fig. 14). This study extended the regression model established using samples collected from 2016 to 2020 through an adaptive sampling strategy to other years, thereby generating annual winter wheat proportion maps for China from 2001 to 2020. In the main winter wheat districts (i.e., 24 provinces) of China, the values of the R2 measured from our results are above 0.9, indicating the obtained winter wheat areas for 2016 to 2020 are highly consistent with the ChinaWheat30L data at the subnational scale (Fig.14a). It is worth noting that the model performance using samples derived from 2016-2020 demonstrates significant temporal transferability, allowing it to be effectively applied to the years 2001-2015, even though it was not trained on samples from that earlier period (Fig.14b). For example, the mean R2 in both 2001-2015 and 2016-2020 is 0.95 and 0.96, respectively. In addition, this study found the fitted regression line lies above the 1:1 line at high values from 2001 to 2015, suggesting an underestimation of winter wheat area in these regions, whereas the opposite pattern was observed from 2016 to 2020. This phenomenon may be attributable to interannual variations in climatic

conditions and crop management practices, as well as to long-term changes in MODIS sensor performance. First, variations in climate and management can influence winter wheat growth, canopy development, and spectral-temporal traits, thereby in turn affecting the estimation of high winter wheat proportions and partly illustrating the underestimation from 2001 to 2015 (Wu et al., 2022). On the other hand, the long-term operational changes in MODIS radiometric calibration and sensor response might cause temporal inconsistencies in spectral observations, which can influence regression estimates and add to the observed bias (Heck et al., 2019).” (page 18, line 376-392)

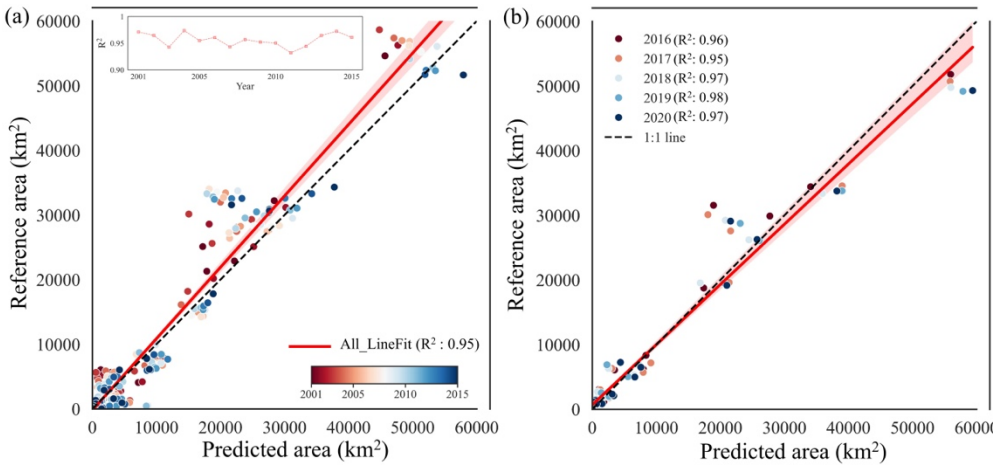


Fig 14. Comparison of region-scale winter wheat areas in China between our results and the ChinaWheat30L data for 2001-2015 (a) and 2016-2020 (b). The line chart in (a) shows the changes in R^2 values for China from 2001 to 2015.

“Temporal changes in winter wheat's spectral and phenological traits, along with MODIS observations, are crucial for long-term monitoring over two decades. Climatic and agronomic changes (e.g., climate warming, cultivar shifts, and changes in sowing dates and agricultural management) can alter winter wheat phenology, thereby affecting its spectral-temporal characteristics over time (Tao et al., 2022). These temporal variations can limit the applicability of winter wheat mapping models across different time periods, particularly in years lacking matching reference samples (Pham et al., 2024; Mohammadi et al., 2024; Lei et al., 2024). Hence, we further evaluated temporal transferability at the subnational scale in China, where the model trained on samples from 2016-2020 achieved mean R² values of 0.95 for 2001-2015 and 0.96 for 2016-2020 (Fig. 14). The comparable R² values between these two periods indicate strong temporal transferability, although our results show a relatively large systematic underestimation from 2001 to 2015. In addition, the extended use of Terra MODIS can lead to uncertainties related to sensor performance, due to instrument response decline and calibration issues. Although ongoing calibration and Collection 6.1 processing greatly enhance the temporal consistency of MODIS observations, previous studies suggest that sensor performance and product stability remain important considerations (Yan et al., 2021; Román et al., 2024).” (page 19-20, line 413-425)

References:

Heck, E., de Beurs, K. M., Owsley, B. C., and Henebry, G. M.: Evaluation of the MODIS collections 5 and 6 for change analysis of vegetation and land surface temperature dynamics in north and South

America, ISPRS J. Photogramm. Remote Sens., 156, 121–134, https://doi.org/10.1016/j.isprsjprs.2019.07.011, 2019.
Lei, L., Wang, X., Zhang, L., Hu, X., and Zhong, Y.: CROPUP: historical products are all you need? An end-to-end cross-year crop map updating framework without the need for in situ samples, Remote Sens. Environ., 315, 114430, https://doi.org/10.1016/j.rse.2024.114430, 2024.
Lobert, F., Löw, J., Schwieder, M., Gocht, A., Schlund, M., Hostert, P., and Erasmi, S.: A deep learning approach for deriving winter wheat phenology from optical and SAR time series at field level, Remote Sens. Environ., 298, 113800, https://doi.org/10.1016/j.rse.2023.113800, 2023.
Mohammadi, S., Belgiu, M., and Stein, A.: A source-free unsupervised domain adaptation method for cross-regional and cross-time crop mapping from satellite image time series, Remote Sens. Environ., 314, 114385, https://doi.org/10.1016/j.rse.2024.114385, 2024.
Pham, V.-D., Tetteh, G., Thiel, F., Erasmi, S., Schwieder, M., Frantz, D., and van der Linden, S.: Temporally transferable crop mapping with temporal encoding and deep learning augmentations, Int. J. Appl. Earth Obs. Geoinf., 129, 103867, https://doi.org/10.1016/j.jag.2024.103867, 2024.
Román, M. O., Justice, C., Paynter, I., Boucher, P. B., Devadiga, S., Endsley, A., Erb, A., Friedl, M., Gao, H., and Giglio, L.: Continuity between NASA MODIS collection 6.1 and VIIRS collection 2 land products, Remote Sens. Environ., 302, 113963, https://doi.org/10.1016/j.rse.2023.113963, 2024.
Tao, F., Zhang, L., Zhang, Z., and Chen, Y.: Climate warming outweighed agricultural managements in affecting wheat phenology across china during 1981–2018, Agric. For. Meteorol., 316, 108865, https://doi.org/10.1016/j.agrformet.2022.108865, 2022.
Yan, K., Pu, J., Park, T., Xu, B., Zeng, Y., Yan, G., Weiss, M., Knyazikhin, Y., and Myneni, R. B.: Performance stability of the MODIS and VIIRS LAI algorithms inferred from analysis of long time series of products, Remote Sens. Environ., 260, 112438, https://doi.org/10.1016/j.rse.2021.112438, 2021.

Comment #1-4. The Adaptive Threshold Concept Is Poorly Defined

The title promises adaptive thresholds, and the GMM-derived optimal threshold is the claimed novelty. Yet:

- The threshold is never formally defined in mathematical terms. What exactly is being thresholded? The RF-predicted wheat fraction? After thresholding, is the output binary or still fractional?

Response: thank you. We revised Section 3.2 to better explain the definition and use of the adaptive threshold. First, we incorporated a mathematical expression to formally specify the thresholding process. Second, we clarified that the thresholding targets the RF-predicted winter wheat fraction at the 1 km pixel level, with a locally adaptive threshold estimated independently for each 100 km grid. Third, we clarified that the threshold output remains a continuous winter wheat fraction rather than a binary wheat/non-wheat classification.

"Specifically, we employed the locally adaptive threshold within each 100 km grid to determine the RF-predicted winter wheat proportion. The thresholding process is defined as follows.

$$F_{i,g} = \begin{cases} \hat{F}_{i,g}, & \hat{F}_{i,g} \geq T_g \\ NoData, & \hat{F}_{i,g} < T_g \end{cases} \quad (4)$$

Where $\hat{F}_{i,g}$ represents the RF-predicted winter wheat proportion for the i-th pixel in the g-th 100 km grid, and T_g is the adaptive threshold for that grid. Pixels with predicted fractions below the

threshold are masked as NoData, whereas those at or above the threshold retain their original predicted proportion values." (page 8-9, line 213-219)

- What does locally optimal mean? Optimal with respect to what objective function? The ROC balanced point is mentioned (L241), but its relationship to the GMM parameters is unclear. The paper states that the GMM parameters are used as inputs to a second Random Forest model that predicts the optimal threshold. This creates a convoluted chain: RF1(MODIS → fractions) → GMM (fractions → parameters) → RF2 (parameters →threshold). The propagation of errors through this chain is never analyzed.

Response: thank you. We are sorry for the unclear terminology for "locally optimal threshold" and "locally adaptive threshold" in the original manuscript. First, we revised the terminology to consistently refer to the threshold estimated from Gaussian features as "locally adaptive threshold" and the threshold derived from the ROC curve as "ROC-derived optimal threshold" in Section 3.2. In addition, the ROC-derived threshold functions solely as a masking step to exclude uncertain winter wheat pixels, and its masked fraction data are subsequently used for GMM fitting.

Due to the complex, non-linear coupling among cloud and atmospheric contamination, cropland mask inaccuracies, and sensor spatial resolution, a fully decoupled, voxel-by-voxel variance decomposition to precisely quantify the isolated contribution of each error source remains unfeasible. Moreover, we acknowledge that our two-stage pipeline carries an inherent risk of error accumulation, where uncertainties in Stage 1 fraction predictions inevitably propagate into Stage 2 GMM curve fitting and threshold selection. However, explicitly disentangling and quantifying these individual error contributions and their exact step-by-step propagation remain challenges.

"Third, due to the complex non-linear coupling among atmospheric contamination, cropland mask inaccuracies, and sensor spatial resolution, a fully decoupled variance decomposition to isolate individual error sources remains unfeasible. Furthermore, our two-stage pipeline carries an inherent risk of error propagation where Stage 1 fraction prediction uncertainties inevitably flow into Stage 2 GMM parameterization and thresholding, making step-by-step error tracking an ongoing methodological limitation." (page 19, line 406-410)

- Figure 4 is referenced but the underlying mathematics are absent. The paper needs equations defining (1) the single and bi-Gaussian models, (2) the ROC threshold derivation, (3) the RF model for threshold prediction, and (4) how the final map is produced from the threshold.

Response: thank you. We added the corresponding mathematics in Section 3.2. We offer precise mathematical definitions for the ROC-derived optimal threshold, the single- and bi-Gaussian models, the RF-based prediction of the locally adaptive threshold, and the final thresholding process used to generate the winter wheat fraction maps.

"We evaluated candidate winter wheat fraction thresholds within each 100 km grid and selected the value that maximized the difference between TPR and the FPR as the ROC-derived optimal threshold.

$$T_g^{roc} = \text{argmax}_t[TPR_g(t) - FPR_g(t)] \quad (1)$$

where T_g^{roc} is the ROC-derived optimal threshold for grid g, and TPR and FPR are the true-positive and false-positive rates at candidate threshold t." (page 8, line 183-187)

"The single and bi-Gaussian models are defined as follows.

$$G(x) = A\exp[-(x-\mu)^2/(2\sigma^2)] \quad (2)$$

where x is the winter wheat fraction, and μ, A, and σ denote the mean, amplitude, and standard deviation, respectively.

$$G(x) = A_1\exp[-(x-\mu_1)^2/(2\sigma_1^2)] + A_2\exp[-(x-\mu_2)^2/(2\sigma_2^2)] \quad (3)$$

where (μ_1, A_1, σ_1) and (μ_2, A_2, σ_2) describe the two Gaussian components within a 100 km grid." (page 8, line 198-203)

"We used the Gaussian parameter features to predict the locally adaptive threshold for each 100 km grid using the random forest regression model.

$$T_g = f_{RF}(\mathbf{Z}_g) = 1/M \sum_{m=1}^{M} h_m(\mathbf{Z}_g) \quad (4)$$

where T_g is the predicted locally adaptive threshold for the g-th 100 km grid, Zg represents the Gaussian parameter features, M is the number of trees in the RF model, and hm denotes the prediction of the m-th regression tree. Specifically, we employed the locally adaptive threshold within each 100 km grid to determine the RF-predicted winter wheat proportion. The thresholding process is defined as follows.

$$F_{i,g} = \begin{cases} \hat{F}_{i,g}, & \hat{F}_{i,g} \geq T_g \\ \text{NoData}, & \hat{F}_{i,g} < T_g \end{cases} \quad (5)$$

Where $\hat{F}_{i,g}$ represents the RF-predicted winter wheat proportion for the i-th pixel in the g-th 100 km grid, and T_g is the adaptive threshold for that grid. Pixels with predicted fractions below the threshold are masked as NoData, whereas those at or above the threshold retain their original predicted proportion values." (page 8-9, line 208-219)

- The choice of 100 km grid is arbitrary. Why 100 km? A sensitivity analysis comparing results at 50 km, 100 km, and 200 km grids is needed.

Response: thank you. We conducted a sensitivity analysis using 50 km, 100 km, and 200 km grids. The corresponding R^2 values were 0.902, 0.972, and 0.983, respectively. Although the 200 km grid achieved the highest R^2, the improvement over the 100 km grid was limited, and the coarser scale could reduce the local representativeness of Gaussian-based thresholds. Hence, we selected the 100 km grid as an appropriate scale for threshold estimation.

"First, we conducted a sensitivity analysis to determine an appropriate grid size by comparing the R^2 values of winter wheat areas from predicted and reference fraction maps for 50, 100, and 200 km grids. The R^2 values of 0.902, 0.972, and 0.983 for the 50 km, 100 km, and 200 km grids, suggesting the mapping agreement improved substantially from 50 to 100 km, whereas increasing the grid size from 100 to 200 km provides only limited improvement. Although the 200 km grid achieved the highest R^2, its coarse scale might merge diverse cropping patterns and weaken the local accuracy of the Gaussian-based thresholds (Xie et al., 2025)." (page 13, line 292-297)

References:

Xie, Y., Nhu, A. N., Song, X.-P., Jia, X., Skakun, S., Li, H., and Wang, Z.: Accounting for spatial variability with geo-aware random forest: a case study for US major crop mapping, Remote Sens. Environ., 319, 114585, https://doi.org/10.1016/j.rse.2024.114585, 2025.

Comment #1-5. Insufficient Error and Uncertainty Analysis

Section 4.4 on uncertainty is only two paragraphs and is far too superficial for a data paper in ESSD:

- No pixel-level uncertainty map is provided. Users need to know where the product is reliable and where it is not. This is standard practice for gridded data products (see e.g., MODIS quality assurance layers, ESA CCI uncertainty maps).

Response: thank you. Following your recommendation, we have generated pixel-level uncertainty maps in 1km resolution across our major validation regions (China, Europe, and the United States). The result indicated that the regional uncertainties remain well within an acceptable threshold; for instance, the mean uncertainty across China is approximately 10%. Moreover, spatial analysis of these maps reveals that higher uncertainties are predominantly localized in peripheral agricultural areas, which are typically characterized by sparse crop distribution and low planting density.

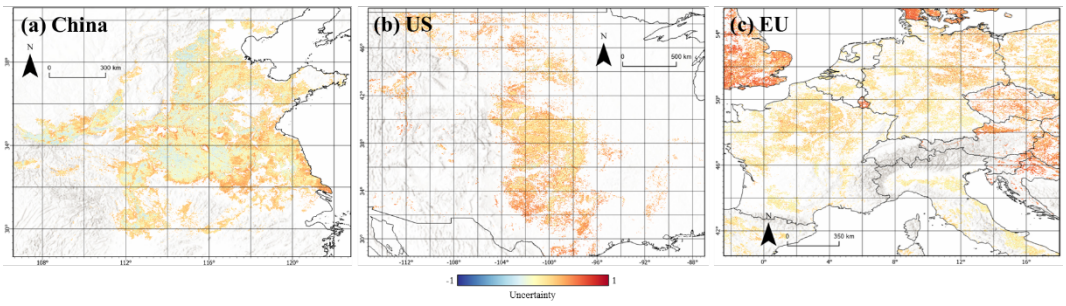


Fig. R1. The uncertainty map in China (a), US (b), and EU (c).

- No breakdown of error sources by magnitude. The authors mention several error sources qualitatively (cropland mask errors, reference data errors, MODIS limitations, systematic RF bias), but never quantify their relative contributions.

Response: thank you. Due to the complex, non-linear coupling among cloud and atmospheric contamination, cropland mask inaccuracies, reference data uncertainties, and sensor spatial resolution, a fully decoupled, voxel-by-voxel variance decomposition to precisely quantify the isolated relative contribution of each individual error source remains unfeasible across global scales. We have now explicitly highlighted this methodological limitation in Uncertainty while expanding our qualitative and semi-quantitative discussion on how these coupled factors interact under different land-cover conditions.

- The systematic regression bias (underestimation of high fractions, overestimation of low fractions) is acknowledged in Section 4.1 but never quantitatively corrected or characterized.

Response: thank you. We clarify that the locally adaptive GMM optimal thresholding step in our pipeline was specifically designed to quantitatively correct and mitigate the inherent

regression bias. Specifically, we effectively removed false-positive low-fraction background noise and rectified the low-end systematic overestimation by adaptive thresholding.

"We finally estimated the locally adaptive threshold for winter wheat proportion using these gaussian features as inputs for each grid, under which value areas are relatively consistent between referenced and predicted fractional maps (Fig.4c)." (page 8, line 207-208)

"This local adaptive thresholding strategy serves primarily as a spatial background filter to prune spurious, low-fraction noise induced by spectral mixing in non-wheat pixels. Furthermore, by accounting for global heterogeneity in field patch sizes, it avoids the systemic overestimation in fragmented landscapes or underestimation in high-density belts that a uniform global threshold would produce." (page 9, line 222-225)

- How do errors propagate through the two-stage RF → GMM → RF pipeline? This is never addressed.

Response: thank you. We acknowledge that our two-stage workflow (i.e., RF → GMM → RF) carries an inherent risk of error accumulation, where initial prediction errors in Stage 1 fraction regression inevitably propagate into Stage 2 GMM curve fitting and subsequent threshold selection. While explicitly disentangling and tracking this exact step-by-step error propagation quantitatively remains a methodological challenge, we have now explicitly incorporated this error propagation mechanism as a limitation in Section 4.5 (Uncertainty) and identified end-to-end uncertainty modeling as a key focus for future work.

"Third, due to the complex non-linear coupling among atmospheric contamination, cropland mask inaccuracies, and sensor spatial resolution, a fully decoupled variance decomposition to isolate individual error sources remains unfeasible. Furthermore, our two-stage pipeline carries an inherent risk of error propagation where Stage 1 fraction prediction uncertainties inevitably flow into Stage 2 GMM parameterization and thresholding, making step-by-step error tracking an ongoing methodological limitation." (page 19, line 406-410)

Comment #1-6. Inadequate Comparison with Existing Products

- No comparison with other global crop maps such as GFSAD (30 m), GCROP (1 km), SPAM (10 km), or the ESA WorldCereal products. These are the products users would compare against when deciding whether to use this dataset.

Response: thank you. We agree that comparison with existing global crop products is useful for assessing dataset applicability. However, these datasets such as GFSAD (i.e., cropland extent), GCROP (i.e., global cropland maps), SPAM (i.e., gridded crop production and harvested area maps), and ESA WorldCereal (i.e., Cereal maps) don't provide consistent global winter wheat classification or proportion products at pixel scale. In addition, we utilized an extensive time series of ChinaWheat30L data to demonstrate our product's good performance over time span, as shown in Comment #1-3.

- No comparison with the MODIS Land Cover product (MCD12Q1) cropland classes, which would be the most natural benchmark given the shared sensor.

Response: thank you. We chose the GLAD cropland dataset at 30 m resolution to more accurately

delineate cropland extent and fragmented fields than the 500 m MCD12Q1 product. Furthermore, the 500 m land-cover classification could lead to more classification errors and uncertainty due to mixed pixels.

"It is worth noting that we utilized the 30 m GLAD cropland dataset rather than the 500 m MCD12Q1 cropland classes to more accurately delineate fragmented cropland and minimize mixed-pixel effects." (page 7, line 179-180)

- The comparison with GlobalWheatYield4km (Section 4.3.2) is ambiguous. The paper claims improved performance (L393) but provides no quantitative basis for this statement. If GlobalWheatYield4km was used for training, comparing against it and claiming improvement is circular.

Response: thank you. We eliminated the GlobalWheatYield4km comparison because the analysis lacked sufficient quantitative evidence to support our assessment of the model's performance. In addition, this dataset was used in sample generation, which compromises the independence of the comparison. Hence, we removed the corresponding analysis and discussion from the revised manuscript. In addition, we added an independent temporal evaluation using the long-term ChinaWheat30L dataset to further assess the performance of our product over an extended time span.

Comment #1-7. Conceptual Gap: Fraction vs. Area

The paper repeatedly conflates winter wheat fraction with winter wheat area. A pixel with 50% wheat fraction means 50% of the 1 km² pixel is planted with winter wheat. When the authors aggregate these fractions to estimate national wheat area (and compare to FAO statistics), they implicitly assume that the fraction values are unbiased and additive. This assumption is never tested. Systematic biases in fraction estimation (acknowledged in Section 4.1) will propagate linearly to area estimates.

Response: thank you. We agree that directly aggregating predicted pixel fractions to regional or national totals implicitly assumes unbiased and additive fraction values, and that systematic bias in fraction estimation can propagate linearly to area estimates. While our spatial aggregation results align with FAO country-level statistics, we acknowledge that fractional bias propagation remains a source of uncertainty. We have added an explicit note in Uncertainty to acknowledge this limitation.

"Finally, we acknowledge that aggregating pixel-level fractions to estimate total wheat area assumes additive and unbiased fraction predictions; thus, any systematic bias in local fraction estimation can propagate linearly into regional and national area totals." (page 19, line 410-412)

Specific Technical Issues

Methodology (Section 3)

Comment #1-8. Section 3.1, Generating regressed fraction maps:

- The adaptive strategies approach (Wen et al., 2022) is referenced at L221 but the method is not described. A reader should not need to consult an external paper to understand the core training procedure.

Response: thank you. We clarify the core training procedure in Section 3.1. Specifically, we

included representative winter wheat samples for years when the model performed relatively poorly. Then, we iteratively retrained the RF model using the updated training dataset to improve temporal robustness.

"It is worth noting that this is an iterative process of training the model for robust performance. We collected representative winter wheat samples and incorporated them into the training dataset for years with relatively poor model performance. We retrained the RF model with the revised samples and iteratively expanded the training dataset to enhance its robustness across years." (page 7, line 166-169)

- How many training samples were collected per country? Per fraction stratum? The stratified sampling is described qualitatively but no sample counts are provided.

Response: thank you. We divided the winter wheat fraction range into 10% intervals and randomly selected the same number of samples from each interval. Approximately 10000 samples were collected for most countries, while about 5000 samples were used for smaller countries with limited spatial extent.

"Specifically, we divided each regressive map from 0 to 100% using the interval of 10%, based on which an equal number of samples were randomly extracted within each category. We collected approximately 10000 samples in total for most countries, while several smaller countries (e.g., Monaco) provided around 5000 samples due to their limited spatial extent." (page 6, line 161-164)

- The MODIS bands used and the temporal compositing window are not specified with sufficient precision. Which 8-day or 16-day composites are used? How are cloudy observations handled?

Response: thank you. We used daily MOD09GA observations and masked cloud, cloud shadow, aerosol, and snow contamination using the pixel QA layer, rather than using 8- or 16-day composite data. We then calculated the monthly 5th, 50th, and 95th percentiles for each spectral band from the remaining valid observations as the spectral-temporal features.

"In addition, we identified and masked cloud, cloud shadow, aerosol, and snow observations by using the pixel quality assurance (QA) layer to retain only valid observations. We computed the 5th, 50th, and 95th percentiles of valid observations for each spectral band monthly, using these as the spectral-temporal features." (page 4, line 115-117)

Comment #1-9. Section 3.2, Gaussian mixture parameters:

- No equations are provided for the GMM. At minimum, Eq. (1) should define the single-Gaussian model and Eq. (2) the bi-Gaussian model.

Response: thank you. We added the corresponding equations for the single-Gaussian and bi-Gaussian models in Section 3.2.

"The single and bi-Gaussian models are defined as follows.

$$G(x) = Aexp[-(x-\mu)^2/(2\sigma^2)] \quad (2)$$

where x is the winter wheat fraction, and μ, A, and σ denote the mean, amplitude, and standard deviation, respectively.

$$G(x) = A_1 exp[-(x-\mu_1)^2/(2\sigma_1^2)] + A_2 exp[-(x-\mu_2)^2/(2\sigma_2^2)] \quad (3)$$

where (μ_1, A_1, σ_1) and (μ_2, A_2, σ_2) describe the two Gaussian components within a 100 km grid." (page 8, line 198-203)

- How is model selection performed (single vs. bi-Gaussian)? AIC? BIC? Visual inspection?

Response: thank you. We did not choose between the single- and bi-Gaussian models based on AIC, BIC, or visual examination. Instead, we used a two-component parameter structure for all grids. We assigned the same Gaussian parameters to both components to ensure consistent inputs for threshold regression on grids with a single Gaussian distribution.

"We applied the single and bi-Gaussian distributions to demonstrate variations in winter wheat planting structure within each grid, rather than to identify specific crop types. It is worth noting that we used two Gaussian components for each grid and applied identical parameters to them in the single Gaussian distribution to maintain consistency for threshold regression parameters." (page 8, line 203-206)

- The ROC mask threshold of 0.206 (Fig. 4a) is cited but its derivation is not explained. Is this value applied globally or per grid?

Response: thank you. We offer precise mathematical definitions for the ROC-derived optimal threshold in Section 3.2. In addition, we applied this value *within each 100 km grid*.

"We evaluated candidate winter wheat fraction thresholds within each 100 km grid and selected the value that maximized the difference between TPR and the FPR as the ROC-derived optimal threshold.

$$T_g^{roc} = \text{argmax}_t[TPR_g(t) - FPR_g(t)] \quad (1)$$

where T_g^{roc} is the ROC-derived optimal threshold for grid g, and TPR and FPR are the true-positive and false-positive rates at candidate threshold t." (page 8, line 183-187)

Comment #1-10. Section 3.3, Evaluation:

- The validation is performed at 5 km aggregation (L422). Why 5 km? How sensitive are the reported metrics to the aggregation scale?

Response: thank you for this comment. **First,** due to widespread crop rotation, fine-resolution pixels (e.g., 10m or 30m) exhibit high interannual variability in crop presence. Here, we spatially aggregated the data to 5 km to stabilize the per-pixel winter wheat proportion and reduce its interannual variation across most regions (Zhang et al., 2018). **Second,** 5 km is widely adopted as a benchmark scale in regional and global Earth Observation-based crop yield forecasting systems (e.g., GLAM) (Becker-Reshef et al., 2018). **Third,** vegetation growth and crop yield models heavily rely on key agro-meteorological drivers (e.g., temperature and precipitation). Major global or regional gridded weather products (such as ERA5-Land, CHIRPS, and NCEP/NCAR reanalysis data) operate at grid resolutions around 5km (Vannoppen and Gobin, 2021). **Finally,** our proposed dataset exhibits low sensitivity to aggregation. Specifically, in China during 2018, the evaluation metrics remained consistently high/stable across spatial scales, with R values of 0.87, 0.77, and 0.86, and RMSE values of

0.09, 0.03, and 0.05 at 1km, 3km, and 5km resolutions, respectively (Fig. Rx.). We have added a justification for choosing 5km in the revised manuscript.

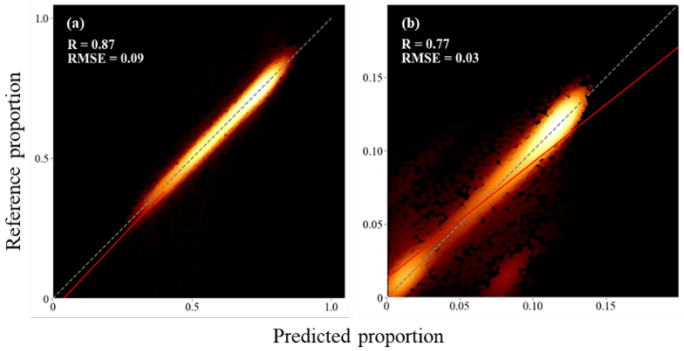


Fig. R2. Comparison of aggregated winter wheat fraction data at 1km (a) and 3km (b) between our results and the other winter wheat products in 2018.

"Furthermore, pixel-level evaluation was conducted at a 5km resolution to stabilize interannual variations from crop rotation, maintain direct spatial compatibility with gridded meteorological drivers and crop yield models (e.g., GLAM) (Becker-Reshef et al., 2018), and ensure scale robustness" (page 10, line 238-241)

Reference:

Becker-Reshef, I., Franch, B., Barker, B., Murphy, E., Santamaria-Artigas, A., Humber, M., Skakun, S., and Vermote, E.: Prior season crop type masks for winter wheat yield forecasting: a US case study, Remote Sens., 10, 1659, https://doi.org/10.3390/rs10101659, 2018.

Vannoppen, A. and Gobin, A.: Estimating farm wheat yields from NDVI and meteorological data, Agronomy, 11, 946–958, https://doi.org/10.3390/agronomy11050946, 2021.

Zhang, X., Xu, H., Jiang, L., Zhao, J., Zuo, W., Qiu, X., Tian, Y., Cao, W., and Zhu, Y.: Selection of appropriate spatial resolution for the meteorological data for regional winter wheat potential productivity simulation in China based on WheatGrow model, Agronomy, 8, 198–215, https://doi.org/10.3390/agronomy8100198, 2018.

Results (Section 4)

Comment #1-11. Section 4.1: The histogram comparison in Fig. 5 uses only 5,000 samples. How representative are these of the full dataset? What is the K-S statistic or Wasserstein distance between true and predicted fraction distributions?

Response: thank you. We clarify the representativeness of the samples and provide a quantitative analysis between the true and predicted fraction distributions in Section 4.1. First, we used stratified random sampling across 10% fraction intervals from 0 to 100% to collect 5,000 samples in these four countries, thereby ensuring coverage across the full fraction range. On the other hand,, we incorporated the Wasserstein distance as a quantitative metric to evaluate the agreement between the true and predicted fraction distributions. The Wasserstein distances in China, the US, France, and Germany are 5.81, 7.85, 12.74, and 15.43, respectively, reflecting a relatively high agreement between the predicted and actual winter wheat fraction distributions.

"We visualized the distribution of 5000 samples with fraction values in the form of frequency

histogram in each country, which were derived from aggregated and regressive maps of winter wheat using stratified random sampling approach from 0 to 100% using the interval of 10% to improve sample representativeness, respectively." (page 11, line 261-263)

"In addition, we further quantified the Wasserstein distance between the true and predicted proportions, which represent the difference in their probability distributions (Naumann, Philip et al., 2026). Overall, these four countries have a low Wasserstein distance, suggesting the predicted proportions in those regions agree well with the true proportions. For example, the Wasserstein distance in China, the US, France, and Germany are 5.81, 7.85, 12.74, and 15.43, respectively." (page 11-12, line 272-276)

References:

Naumann, P., Kauffmann, J., and Montavon, G.: Wasserstein distances made explainable: insights into dataset shifts and transport phenomena, IEEE Trans. Pattern Anal. Mach. Intell., https://doi.org/10.1109/TPAMI.2026.3656947, 2026.

Comment #1-12. Section 4.3.1: The R^2 values in Fig. 7 pool all countries and years. This masks interannual and inter-country variability. The paper should report per-country and per-year metrics.

Response: thank you. The FAO data provides only one national winter wheat area value for each country in a given year, so we cannot calculate an R^2 value separately for each country-year combination. We added an independent comparison with the ChinaWheat30L dataset to provide a more detailed temporal evaluation in **Comment #1-3**. The model trained on data from 2016-2020 achieved mean R^2 values of 0.95 for 2001-2015 and 0.96 for 2016–2020, suggesting it has a strong ability to transfer across these time periods.

Comment #1-13. The two distinct clusters in China's scatter plots (L428–429) are attributed to higher planting density in China. This is a post-hoc explanation; the authors should test this hypothesis by comparing the high-fraction and low-fraction clusters against independent data on planting density or field size.

Response: thank you. Rather than a speculative post-hoc assumption, this distinct bimodal pattern is a direct reflection of China's unique dual-patterned agricultural structure, which contrasts sharply with Europe and the US. Specifically, corresponds to major intensive agricultural basins (e.g., the North China Plain), where winter wheat is planted in large, continuous, and highly concentrated field blocks (Fig. S5a). Moreover, in mountainous terrain or transitional farming zones (e.g., parts of Southern China), where smallholder farming predominates, fragmented and dispersed wheat patches are common (Fig. S5b). In contrast, winter wheat in the US is extensively intercropped or rotated in large-scale mosaic patterns alongside other crops, generating a broad, continuous gradient of mix-fraction pixels (Fig. S5c). European wheat landscapes are characterized predominantly by medium-to-small contiguous parcels surrounded by fixed natural borders (forests, pastures), leading to a more localized or unimodal cluster distribution (Fig. S5d).

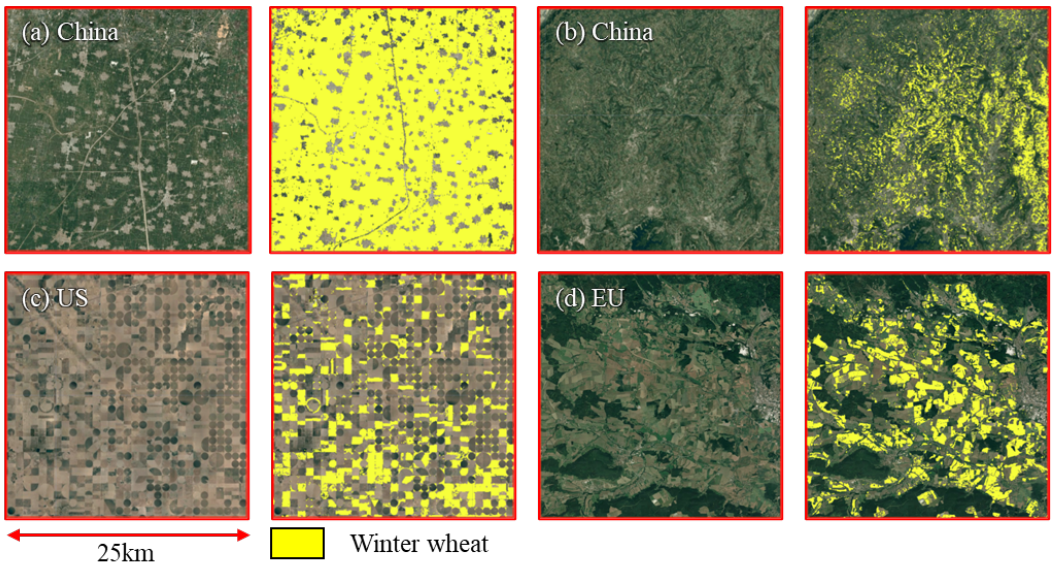


Fig. S5. Comparison of the agricultural structure of winter wheat in China (a and b), the US (c), and the EU (d).

"This distinct bimodal pattern in China reflects a dual agricultural structure, where high-fraction clusters correspond to large, contiguous wheat blocks in intensive agricultural basins (e.g., the North China Plain), whereas low-fraction clusters correspond to fragmented, smallholder patches in mountainous or transitional zones (Fig. S5). In contrast, intercropped crop rotation mosaics in the US generate a continuous fraction gradient, while European landscapes dominated by bordered, medium-to-small contiguous parcels produce a more localized cluster distribution." (page 16-17, line 349-353)

Figures

Comment #1-14. Several figures are described only in captions without sufficient standalone interpretability (e.g., Fig. 6, Fig. 10).

Response: thank you. We provide sufficient information for standalone interpretation. First, we clarified the comparison of winter wheat areas derived from the reference and predicted fractions at the 100 km grid scale in China, the US, and EU countries, and defined the boxplots, violin plots, mean values, and median values in Fig. 7 of the revised manuscript. In addition, as noted in Comment #1-6, we eliminated the GlobalWheatYield4km comparison because the analysis lacked sufficient quantitative evidence to support our assessment of the model's performance. Therefore, both Fig. 10 and the six-site analysis were removed from the revised manuscript.

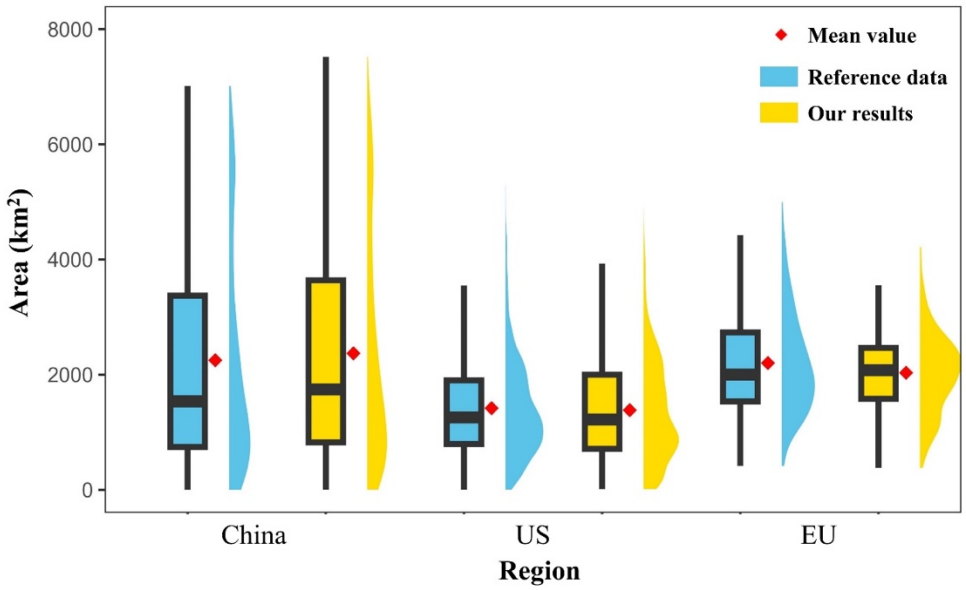


Fig. 7. Comparison of the derived areas using gaussian parameter derived optimal thresholds at the 100 km grid scale in China, the US, and EU countries. Blue and yellow represent the reference data and our results, respectively. The boxplots and violin plots show the area distributions. Red diamonds indicate the mean values, and black bold lines denote the median values.

Comment #1-15. Fig. 10 claims six sites but the text (L402) references them with symbols that do not render clearly in the manuscript.

Response: thank you. As noted in Comment #1-6, we eliminated the GlobalWheatYield4km comparison because the analysis lacked sufficient quantitative evidence to support our assessment of the model's performance. In addition, this dataset played a role in sample generation, which compromises the independence of the comparison. Therefore, both Fig. 10 and the six-site analysis were removed from the revised manuscript.

Minor Issues

Comment #1-16. L150: 500 km should be 500 m — this typo could confuse readers.

Response: thank you. Done.

Comment #1-17. L239: The GLAD cropland dataset reference (Potapov et al., 2022) should specify the version and year used.

Response: thank you. We used all five available GLAD global cropland extent layers for 2003, 2007, 2011, 2015, and 2019 at 30 m resolution and selected the temporally nearest layer for each mapping year during cropland masking.

"Besides, we retained the cropland extent using the five available GLAD global cropland extent layers (i.e., 2003, 2007, 2011, 2015, and 2019) at 30 m resolution and selected the temporally nearest cropland layer for each mapping year to mask the non-cropland areas to reduce uncertainties (Potapov et al., 2022)." (page 7, line 176-179)

References:

Potapov, P., Turubanova, S., Hansen, M. C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., and Cortez, J.: Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century, Nat. Food, 3, 19–28, https://doi.org/10.1038/s43016-021-00429-z, 2022.

Comment #1-18. Abstract: States $R^2 = 0.81$ with FAO statistics, but Section 4.3.1 does not list this exact number in context. The abstract should specify whether this is the mean, median, or overall R^2.

Response: thank you. The value of 0.81 is the minimum annual R^2 over the 20-year period (2001–2020), not a mean, median, or overall R^2. We revised the corresponding description to ensure consistency between the Abstract and the Results section.

"Results demonstrated that our mapping approach aligns closely with existing local winter wheat maps and statistical data, achieving annual coefficients of determination (R^2) above 0.81 with FAO statistics in primary planting regions and exceeding 0.72 at subnational scales." (page 1, line 28-30)

"The R^2 values, all above 0.81, confirm that the mapped winter wheat areas are in strong agreement with FAO statistics at the national scale." (page 14, line 308-309)

Comment #1-19. L82: coarse-resolution satellite observations (e.g., 250 m, 500 m, and 1000 m) — 250 m and 500 m are not coarse for crop mapping; this statement is misleading. The issue is mixed pixels, not coarseness perse.

Response: thank you. We removed the inappropriate "coarse-resolution" description in the revised manuscript. Then, we clarified that mixed pixels and spectral confusion within crop regions are the primary limitations on crop mapping performance at these spatial resolutions.

"Only limited efforts have been made to map crop dynamics at large spatial scales, particularly using satellite observations at spatial resolutions of 250-1000 m. This is because the disturbance of mixed pixels and spectral confusion within crop regions may affect the mapping performance (Zhang et al., 2021)." (page 2, line 58-60)

Comment #1-20. English language: The manuscript would benefit from careful proofreading. There are several awkward constructions (e.g., robust and harmonious random forest regression model at L206—what does harmonious mean in this context?; demonstrated significant performance at L505).

Response: thank you. We carefully proofread the entire manuscript and revised awkward or unclear expressions. For example, we removed the inappropriate term "harmonious" from the description of the random forest regression model and revised the expression "significant performance" to "good performance" when describing the model's mapping performance.

"We first created the stratified samples of winter wheat fractions at 1km resolution from multiple public crop datasets, and then developed a robust random forest regression model using MODIS surface reflectance." (page 1, line 23-25)

"On the other hand, the proposed fraction mapping method appeared good performance at the regional scale," (page 20, line 437)

Comment #1-21. Reference formatting: Some references lack DOIs or have incomplete information

(e.g., L539, AAFC reference).

Response: thank you. We thoroughly reviewed the reference list to fill in any missing or incomplete details. Specifically, we updated the AAFC, FAOSTAT, and USDA-NASS references by including the complete dataset or webpage titles, the responsible institutions, official data links, and access details where relevant. Additionally, we reviewed the remaining references and included DOI details when available.

"AAFC: Annual crop inventory 2025, https://open.canada.ca/data/en/dataset, 2025., 2025." (page 21, line 457)

"FAOSTAT: Crops and livestock products, https://www.fao.org/faostat/en/#data/QCL, 2025., 2025." (page 22, line 484)

"USDA-NASS: Cropland data layer metadata, https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/meta.php, 2025., 2025." (page 25, line 600-601)

Reviewer #2

Comment #2-0. This manuscript presents a novel framework that integrates 20 years of MODIS data with Random Forest (RF) regression and a Gaussian Mixture Model (GMM) to map global winter wheat fractional cover at 1 km resolution. Addressing sub-pixel heterogeneity and mixed-pixel decomposition at a global scale over two decades is highly challenging, and the resulting dataset holds immense potential for crop growth modelling and global yield forecasting. The technical execution shows promise, and the extensive validation effort is commendable. However, in its current form, the manuscript suffers from several methodological gaps and presentation deficiencies that preclude immediate publication. I recommend a revision to address these concerns

Response: Thank you very much for your positive comments. We have carefully revised the manuscript in response to your feedback and provided a detailed, point-by-point response below.

Comment #2-1. This study presented an effective approach to solving global mixed-pixel challenges of winter wheat mapping by integrating the GMM with the RF regression model. However, the current methodological description of the GMM, the "adaptive strategy," and the parameter extraction is overly concise. Please revise the manuscript to include mathematical formulations of the GMM and clear algorithmic workflows that explain step by step how the GMM curves are decomposed and how the adaptive parameters are updated.

Response: thank you. We clarify the core training procedure in Section 3.1. Specifically, we included representative winter wheat samples for years when the model performed relatively poorly. Then, we iteratively retrained the RF model using the updated training dataset to improve temporal robustness. In addition, we added the corresponding mathematics in Section 3.2. We offer precise mathematical definitions for the ROC-derived optimal threshold, the single- and bi-Gaussian models, the RF-based prediction of the locally adaptive threshold, and the final thresholding process used to generate the winter wheat fraction maps.

"It is worth noting that this is an iterative process of training the model for robust performance. We collected representative winter wheat samples and incorporated them into the training dataset for years with relatively poor model performance. We retrained the RF model with the revised samples and iteratively expanded the training dataset to enhance its robustness across years." (page 7, line 166-169)

"We evaluated candidate winter wheat fraction thresholds within each 100 km grid and selected the value that maximized the difference between TPR and the FPR as the ROC-derived optimal threshold.

$$T_g^{roc} = \text{argmax}_t[TPR_g(t) - FPR_g(t)] \quad (1)$$

where T_g^{roc} is the ROC-derived optimal threshold for grid g, and TPR and FPR are the true-positive and false-positive rates at candidate threshold t." (page 8, line 183-187)

"The single and bi-Gaussian models are defined as follows.

$$G(x) = A\exp[-(x-\mu)^2/(2\sigma^2)] \quad (2)$$

where x is the winter wheat fraction, and μ, A, and σ denote the mean, amplitude, and standard deviation, respectively.

$$G(x) = A_1 \exp[-(x-\mu_1)^2/(2\sigma_1^2)] + A_2 \exp[-(x-\mu_2)^2/(2\sigma_2^2)] \quad (3)$$

where (μ_1, A_1, σ_1) and (μ_2, A_2, σ_2) describe the two Gaussian components within a 100 km grid." (page 8, line 198-203)

"We used the Gaussian parameter features to predict the locally adaptive threshold for each 100 km grid using the random forest regression model.

$$T_g = f_{RF}(\mathbf{Z}_g) = 1/M \sum_{m=1}^{M} h_m(\mathbf{Z}_g) \quad (4)$$

where T_g is the predicted locally adaptive threshold for the g-th 100 km grid, Zg represents the Gaussian parameter features, M is the number of trees in the RF model, and hm denotes the prediction of the m-th regression tree. Specifically, we employed the locally adaptive threshold within each 100 km grid to determine the RF-predicted winter wheat proportion. The thresholding process is defined as follows.

$$F_{i,g} = \begin{cases} \hat{F}_{i,g}, & \hat{F}_{i,g} \geq T_g \\ \text{NoData}, & \hat{F}_{i,g} < T_g \end{cases} \quad (5)$$

Where $\hat{F}_{i,g}$ represents the RF-predicted winter wheat proportion for the i-th pixel in the g-th 100 km grid, and T_g is the adaptive threshold for that grid. Pixels with predicted fractions below the threshold are masked as NoData, whereas those at or above the threshold retain their original predicted proportion values." (page 8, line 208-219)

Comment #2-2. The pursuit of an adaptive local thresholding strategy shows a commendable understanding of geographical heterogeneity. Since the ultimate product of this study is a continuous fractional winter wheat map, the need to determine an "optimal threshold" for masking remains unclear. The authors need to explicitly clarify whether this threshold is used merely to exclude low-fraction noise or if it serves a different purpose in the workflow.

Response: thank you. **On one hand,** the MODIS imagery with a resolution of 500m inherently suffers from spectral mixing and regression smoothing effects, which often generate spurious low-fraction predictions in non-wheat areas or non-cropland pixels. Here, we used a local adaptive threshold as a spatial cutoff to filter out background noise pixels while preserving low-density wheat fields. **On the other hand,** cropping structures and field patch sizes vary globally; using a uniform global threshold would lead to overestimation in highly fragmented landscapes or underestimation in high-density agricultural regions. Hence, we derived local optimal thresholds by modeling with GMM parameters to align predicted fractional sums with referenced spatial extents within each 100km grid.

"This local adaptive thresholding strategy serves primarily as a spatial background filter to prune spurious, low-fraction noise induced by spectral mixing in non-wheat pixels. Furthermore, by accounting for global heterogeneity in field patch sizes, it avoids the systemic overestimation in fragmented landscapes or underestimation in high-density belts that a uniform global threshold would produce." (page 9, line 222-225)

Comment #2-3. The validation effort across highly diverse agricultural regions in the US, China, and Europe demonstrates the scientific rigor of this work. In Fig. 5, a distinct systematic error is visible for France and Germany, characterized by a noticeable gap in the low-proportion range (0–

20%), but it is absent in the US and China. Please provide a more detailed geographical and agricultural explanation for this regional disparity, rather than attributing it to general machine-learning regression biases and data quality issues.

Response: thank you. Rather than relying solely on general machine-learning regression biases or data quality issues, this regional disparity stems from landscape structure, land-cover spatial continuity, and regional farming practices.

On one hand, in France and Germany, winter wheat fields in low-proportion pixels often appear as small, isolated patches surrounded by non-crop land cover (e.g., dense grasslands, broadleaf forests, or built-up infrastructure), producing extremely weak signals that are easily submerged by the strong background spectral signal of surrounding vegetation. Specifically, at the coarse 500m MODIS scale, these minor, highly localized plantings fail to generate a strong enough phenological NDVI amplitude to be recognized as low-fraction wheat, resulting in the observed gap in the predicted low-proportion range.

On the other hand, Western European agriculture is characterized by high land-cover spatial continuity within agricultural parcels bounded by sharp, vector-defined landscape edges, exhibiting abrupt transitions between pure agricultural blocks and non-agricultural land. Consequently, 1km pixels covering European landscapes either contain moderate-to-high crop fractions within intensive basins or fall into surrounding non-crop zones. In our pipeline, these weak, low-value signals are easily treated as spatial background noise and pruned by the adaptive thresholding algorithm, resulting in sparse predicted values in the 0–20% interval. We have thoroughly revised Section 4.2 to incorporate this multi-dimensional geographical discussion.

"This phenomenon is attributable to (1) in Western Europe, small and isolated wheat patches generate weak spectral signals surrounded by dense non-cropland features. At coarse resolutions, these weak mixed signatures resemble background noise and are subsequently masked by our local adaptive thresholding algorithm," (page 11, line 265-268)

Comment #2-4. I recommend adding a dedicated table to outline the specific spectral bands and vegetation indices included in the RF regression model.

Response: Thank you. As suggested, we have added a table (Table S1) to the revised manuscript that summarizes the seven surface reflectance bands from MOD09GA v061, the 12-month temporal window spanning the complete winter wheat growth cycle, and the statistical percentiles (e.g., 5th, 50th, and 95th) used to generate monthly composites, resulting in a total of 252 predictors for the Random Forest regression model.

Table S1. The input feature for the Random Forest model.

Source Data	Spectral Bands	Monthly Quantiles	Temporal Sequence
MODIS (MOD09GA.061)	Red NIR Blue Green	5th Percentile 50th Percentile 95th Percentile	12 monthly windows (i.e., spanning the full winter wheat growing season)

	SWIR-1	
	SWIR-2	
	SWIR-3	

"We generated fractional maps of winter wheat using the robust random forest regression model in each country (Fig.3) with 252 predictors (Table S1)." (page 6, line 151-152)

Comment #2-5. Because this is a global-scale study spanning diverse climate zones, using absolute calendar months to align time-series data may introduce errors; the authors should detail how their model handles the spatial phenological heterogeneity.

Response: Thank you. **First,** rather than applying a universal timeline, we adjusted the 12-month processing window to align with crop calendars. Specifically, we set the timeline from September to August of the following year for the Northern Hemisphere, and from April to March of the following year for the Southern Hemisphere to span the complete winter wheat growth cycle.
Second, instead of relying on single-date images or multi-day composite dates, we calculated monthly 5th, 50th, and 95th percentiles of each selected surface reflectance band to provide statistical smoothing, reducing sensitivity to minor phenological shifts (e.g., weekly variations in green-up dates) and cloud contamination.

Finally, we used the Random Forest regression model to learn multi-modal phenological profiles across different agro-ecological zones (i.e, divided by 100km grid), and a locally adaptive thresholding strategy to compensate for residual phenological phase shifts between regional climate zones.

"Rather than applying a universal timeline, we adjusted the 12-month processing window to align with crop calendars—setting the timeline from September to August of the following year for the Northern Hemisphere and from April to March for the Southern Hemisphere to cover the full winter wheat growth cycle." (page 4, line 112-115)

"We computed the 5th, 50th, and 95th percentiles of valid observations for each spectral band monthly, using these as the spectral-temporal features." (page 4, line 116-117)

"That is, we developed the robust regression model of winter wheat identification based on training samples to estimate the locally adaptive thresholds for each local grid, while the remaining samples were used for validating threshold estimation in our study. This local adaptive thresholding strategy serves primarily as a spatial background filter to prune spurious, low-fraction noise induced by spectral mixing in non-wheat pixels. Furthermore, by accounting for global heterogeneity in field patch sizes, it avoids the systemic overestimation in fragmented landscapes or underestimation in high-density belts that a uniform global threshold would produce." (page 9, line 220-225)

Comment #2-6. This fractional mapping framework bridges the gap between temporally inconsistent high-resolution binary maps and coarse-resolution mixed pixels. To maximize the impact of this study, please articulate your product's unique value proposition more clearly in the Introduction and Discussion.

Response: thank you. High-resolution satellite observations (e.g., Landsat, Sentinel-2) provide fine spatial details, but their long return intervals and cloud contamination often cause temporal gaps and spatial inconsistencies in binary crop mapping over multi-decadal timeframes. Conversely, coarse-resolution satellite products provide daily continuous observations, but standard binary classification algorithms force mixed pixels into single land-cover labels, leading to severe misclassification or total exclusion of fragmented/low-density cropland. Our 1 km global fractional dataset bridges this critical methodological gap by providing a long-term (2001–2020), temporally consistent, and sub-pixel continuous representation of winter wheat dynamics. This offers unprecedented sub-pixel fidelity for global crop yield modeling, agro-meteorological assessment, and policy-making, without fine-resolution temporal gaps or coarse-resolution signal truncation. We have updated the Introduction to explicitly articulate this unique contribution.

"Furthermore, our fractional mapping framework bridges the critical gap between temporally inconsistent high-resolution binary maps and coarse-resolution mixed pixels. It delivers a long-term, spatially continuous, and sub-pixel-accurate benchmark for global winter wheat dynamics, maximizing its potential for regional crop yield modelling, climate impact assessment, and agricultural policy decision-making." (page 3, line 95-98)

Comment #2-7. A deep grammatical and stylistic revision is needed throughout the manuscript to correct several typos and awkward expressions. For example, in Section 4.3.3, there is a glaring typographical error: " high 1accuracy..." which should be corrected to "high accuracy."

Response: Thank you. We apologize for these language and typographical oversights. We have performed a grammatical and stylistic revision across the entire manuscript. The specific typo in Section 4.3.3 ("high 1accuracy") as well as other identified minor language issues and awkward expressions have all been corrected.

"The result indicated that our fractional mapping results had high accuracy in four different regions in 2018, with R-value and RMSE as 0.78 and 0.04, respectively (Fig. 12). Moreover, the proposed fraction mapping framework demonstrated high accuracy in producing reliable results across different years (i.e., the mean R-value and RMSE were 0.84 and 0.05 in China from 2016 to 2020, and 0.74 and 0.04 in the US from 2008 to 2020)." (page 16, line 343-346)

Comment #2-8. There are several mismatches between the figures and their corresponding text citations that must be thoroughly audited. For instance, in Section 4.3.2, the text refers to Fig. 11b; however, there is no sub-figure "11b" in the manuscript.

Response: Thank you. We have conducted a comprehensive audit of all figures, sub-figures, and their corresponding callouts in the manuscript. All misalignments—including the incorrect reference to "Fig. 11b" in Section 4.3.2—have been fully corrected and cross-checked to ensure complete consistency throughout the revised manuscript.